Efficient Lexical Encoding of Natural Language

Heers, Brian (School: InTech Collegiate Academy)

This project developed and evaluated NLC (Natural Language Compression), a language-agnostic lexical text encoder optimized for natural language (NL). Efficient NL encoding is most applicable for bandwidth-constrained systems, including satellite text communication, backup/emergency low-bitrate communication systems, decentralized or post-disaster mesh network messaging, remote research, and other applications where reduced bandwidth is necessary for data reception or cost reduction. Unlike standard byte/character-oriented and general-purpose encoders, NLC encoded lexemes as single units. A lexeme was defined as a word, punctuation mark, word combination ("such as"), or word segment (-"ing"). NLC assigned identifiers to lexemes, which were variably encoded with bit-lengths proportional to its frequency in a user's isolect (derived from a corpus shared between encoder and decoder). NLC used contextual sub-lexicons to predict succeeding lexemes, analyzed standard usages of punctuation and spaces to implicitly encode whitespace characters from context, and morphologically segmented unknown words into sub-lexemes. Tests were individually performed on 1,757,861 sentences across seven languages, including English. In English, NLC encoded at a median of 2.47 BPC (Bits Per Character), Q1 2.18, and Q3 2.85. While some existing compression algorithms achieve lower benchmarked BPCs on large corpora, they are computationally intensive or degrade in efficiency on short messages. NLC is a lightweight encoder, suitable for underpowered devices, that would reduce transmission costs for short NL messages in bandwidth-limited environments. Future work will explore optimizing unusual NL usage, language-specific optimizations, and porting to lower-level programming languages.

Awards Won:
Association for the Advancement of Artificial Intelligence: AAAI Student Memberships for each finalist that is part of the 1st, 2nd, and 3rd Prize Winning projects and 5 Honorable Mention winning projects (up to 3 students per project) (in-kind award / part of the 1st-3rd prize)